FINAL PROJECT PROPOSAL

Machine Learning For Network Intrusion Detection

Students:
Hussain Ahmad Mohammad
Jiayi Yao
Michael Fortunato

Department of Computer Science

November 16, 2024

1 Proposal

Network Intrusion Detection Systems (NIDS) promise to detect and report abnormal and potentially malicious traffic on a computer network. Researchers have long sought the potential of statistical classifiers to robustly classify network traffic as benign or malicious, and thus be used to implement a sophisticated NIDS. Of the efforts, the current state of the art NIDS system is Kitsune [1], which classifies a packet $\mathbf{x}$ as *benign* or *malicious* by first learning a feature mapping ψ from $\mathbf{x}$ and to a hidden space, and then learning a classifier from that hidden space to the binary label. By first mapping an observed packet $\mathbf{x}$ into a hidden learned feature space, and then classifying *benign* or *malicious* on this latent space, Kitsune was able to robustly classify packets as being part of an attack and so implement a powerful NIDS.

However, in the years since the initial publication of Kitsune, it has been clearly demonstrated that classifiers which use auto-encoders, such as Kitsune, are quite vulnerable to adversarial attacks, in which the attacker manufactures fake input to fool the classifier [2]. In our work, we will explore how well a network intrusion classifier fairs when using simpler techniques of feature mapping like SVD and kernel regression to classify whether a packet is malicious or benign, as compared to more sophisticated classifiers like Kitsune. In particular, we will use the economy Singular Value Decomposition against our data matrix X to compute the most left singular vectors, and use that to compute our feature mapping function Φ. This is in contrast to Kitsune, which used the method of auto-encoders to learn its neural network based feature mapping function Φ. After that, we will train a classifier on our mapped feature matrix $\mathbf{Z}$. Then we will compare the prediction accuracy of our classifier against Kitsune on a common dataset using a technique such as cross-validation. Afterwards, we may possibly investigate how our model fairs against an adversarial attack as compared to Kitsune, which has recently been shown to be vulnerable to adversarial attacks

Briefly, we define our model formally and discuss the potential datasets we will use to compare it with Kitsune. Given a packet $\mathbf{x} \in \mathbb{R}^p$, which is described by p scalar metrics we have chosen to observe, such at the current average download speed of the network, we wish to classify $\mathbf{x}$ as being part of either *benign* or a *malicious* event. To do this, we first learn a feature mapping $\psi : \mathbf{R}^p \mapsto \mathbf{R}^k$, which maps $\mathbf{x}$ to some learned hidden space. We then learn the function $f : \mathbf{R}^k \mapsto [0, 1]$, and classify $\mathbf{x}$ as being *malicious* if $f(\psi(\mathbf{x})) > b \in [0, 1]$ for some bias b and *benign* otherwise. We will construct

Hussain Ahmad Mohammad
Jiayi Yao
Michael Fortunato

our feature mapping ψ using well-known general techniques such as economy SVD, kernel regression, and support vector machines. This is in contrast to Kitsune's feature map ζ, which is a neural network trained using the technique of auto-encoders. We will then compare the prediction accuracy of our model prediction's against Kitsune's. As further work, we may possibly investigate how our model fairs against an adversarial attack as compared to Kitsune, which has recently been shown to be vulnerable to adversarial attacks [2][3].

There have been efforts to construct high-quality datasets for both training and benchmarking intrusion detection classifiers. As of 2024, there are three datasets which are used as reference benchmarks. They are the Kitsune Network Attack Dataset[4], the CSE-CIC-IDS2018 [5], and the older UNSW-NB15 Dataset [6]. We will briefly describe the Kitsume Network Attack Dataset as it the most popular. The dataset consists of seven sub-datasets, where each sub data set is a time series of data that was recorded when a given attack event was taking place. For explanation purposes, let us consider the SYN Dos sub dataset, which is the time series of data when the network was experiencing an attack. Precisely put, the sub-dataset consists of $n = 1,000,000$ rows and $p = 115$ feature columns. The i-th row of this dataset represents the observation of the network at time step i, as described by the p features. We can arrange the data into a $n \times p$ matrix $\mathbf{X}$. In addition, we are provided with y labels, where $y_i = 1$ if a Dyn DOS attack was occurring at time-step i, and 0 otherwise. The Kitsune dataset is appealing because of it's reputation and the fact that it describes real data, but the process of using the other datasets would be the same.

We will make our findings public and publish our code so that our work may be reproduced and extended upon by the community.

References

[1] Y. Mirsky, T. Doitshman, Y. Elovici, and A. Shabtai, "Kitsune: An ensemble of autoencoders for online network intrusion detection," in *NDSS Symposium 2018*, 2018. [Online]. Available: `https://www.ndss-symposium.org/wp-content/uploads/2018/02/ndss2018_03A-3_Mirsky_paper.pdf`.

[2] J. Costa, F. Apolinário, and C. Ribeiro, "Argan-ids: Adversarial resistant intrusion detection systems using generative adversarial networks," in *Proceedings of the 19th International Conference on Availability, Reliability and Security (ARES 2024)*, Vienna, Austria: ACM, Jul. 2024, p. 10. [Online]. Available: `https://doi.org/`.

[3] A. unknown, "Cascaded multi-class network intrusion detection with decision tree and self-attentive model," in *IEEE Transactions on Emerging Topics in Computing*, 2023. [Online]. Available: `https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10031221`.

[4] Kaggle, *Kitsune network attack dataset*. [Online]. Available: `https://www.kaggle.com/datasets/ymirsky/network-attack-dataset-kitsune`.

[5] C. I. for Cybersecurity, *Cse-cic-ids2018 on aws*, 2018. [Online]. Available: `https://www.unb.ca/cic/datasets/ids-2018.html`.

[6] U. C. Cyber, *The unsw-nb15 dataset*, 2015. [Online]. Available: `https://research.unsw.edu.au/projects/unsw-nb15-dataset`.